REMARKS/ARGUMENTS

Claim 1-18 are currently pending in the present application, with no claims being cancelled at this time. Claims 1, 4, 13, 14, 15, 17 and 18 have been amended in order to better clarify the claimed subject matter. Applicants respectfully submit that no new matter within the meaning of 35 USC §132 has been added.

In light of the above amendments and following remarks, Applicants respectfully submit that all objections and rejections have been overcome, and that the application is now in condition for allowance.

Claim Objections

In Claim 17, the word "organising" has been corrected to "organizing" as requested by the examiner in paragraph 3 of the Office Action. Applicants respectfully request withdrawal of this objection.

Claim Rejections - 35 USC § 112

As requested by the Examiner in paragraph 4, the word "term" in Claim 1 has been clarified by inserting the phrase "each term comprising one or more word". Corresponding amendments have been made to independent claims 13, 14, 15 and 18. Accordingly, Applicants respectfully request withdrawal of this rejection.

Claim Rejections - 35 USC § 101

To address the Examiner's objections at paragraph 6 of the office action, Claims 13 and 18 have been reworded as suggested by the Examiner. Accordingly, Applicants respectfully request withdrawal of this rejection.

Claim Rejections - 35 USC § 103

Applicant agrees with the Examiner that United States Patent Application US 2003/0065635 (Sahami) is not concerned with cluster attractors. Instead of clustering data using attractors, Sahami clusters data using a totally different technique based on the influence of the attributes of the data (see in particular paragraph [0061] and Claim 1, lines 12 to 27 of Sahami, see also paragraphs [009] and [0010] for a description of what is meant by "attribute"). Thus, Sahami neither needs nor uses attractors to create his clusters.

Moreover, Sahami is not concerned with clustering documents comprising at least one term, each term comprising one or more words. Instead, Sahami discloses a method for clustering structured data that can be presented as records of pairs, each pair comprising an "attribute" and an "attribute value" (see for example Sahami's abstract, second sentence and Claim 1, lines 1 to 6 of Sahami). In contrast, Claim 1 relates to identifying cluster attractors for documents comprising unstructured text, e.g. paragraph, sentences, phrases, words, stems of words and so on. The clustering of structured data as contemplated by Sahami is significantly different from the clustering of unstructured information. There is no straightforward method of representing an unstructured document as a set of attributes with specific values as required by Sahami.

Applicant agrees with the examiner that Sahami discloses the use of probability distributions when creating his clusters. However, Sahami does not disclose the calculation of probability distributions that are indicative of the frequency of occurrences of terms in the documents as stipulated by Claim 1. Instead, Sahami uses conditional probability to evaluate the influence of an attribute given a cluster of data records (see paragraphs [0066] and [0068] of Sahami). It is emphasised that Sahami is not calculating probabilities concerning the frequency of occurrence of terms (words) in documents because, as indicated above, Sahami is concerned only with the "attributes" of structured data and not the contents (words) of an unstructured document.

With regard to paragraph [0027] of Sahami, the COBWEB technique is presented as a prior art technique for performing clustering using tree structures and is not actually used by Sahami. In any event, probability is used by the COBWEB technique to assign data points to data clusters. For example, referring to Figure 4 of Sahami, if a new record X equals (x1, x2…xN) has its first attribute value equal to x1, then with probability 1.0, this record should be assigned to cluster C_1. Hence, the probability techniques disclosed by COBWEB do not relate to the frequency of occurrence of terms (words) within documents.

Sahami does not disclose calculating the entropy of the respective probability distributions. As indicated above, Sahami does not disclose the calculation of probability distributions as defined by Claim 1 and so it follows that it cannot disclose the calculation of the entropy of those probability distributions. Sahami does disclose the use of entropy as part of its calculations in paragraphs [0083] to [0087], but only as a means of eliminating features that are not useful in identifying clusters - see in particular paragraphs [0084] to [0086].

It follows from the foregoing that Sahami does not disclose selecting at least one probability distribution as a cluster attractor depending on the respective entropy value.

In summary, Sahami does not disclose the following features of Claim 1:

A. A method of determining cluster attractors. Sahami does not use cluster attractors during his clustering process.
B. Using a plurality of documents, each document comprising at least one term, each term comprising at least one word.

Instead, Sahami's calculations are based on structured data wherein "attributes" are assigned in pairs to the data.

C. Calculating a probability distribution indicative of the frequency of occurrence of the, or each, other term that co-occurs with said term in at least one of said documents. Instead, Sahami uses conditional probability to evaluate the influence of an attribute given a cluster of data records.

D. Calculating, in respect of each term, the entropy of the respective probability distribution. Instead, Sahami only teaches the use of entropy as a means of eliminating unwanted features of the data set.

E. Selecting at least one of said probability distributions as a cluster attractor depending on the respective entropy value. This follows from the comments made at A to D above.

It is respectfully submitted therefore that Claim 1 is novel over Sahami by virtue of each of Features A to E recited above.

With regard to United States Patent US 5,787,422 (Tukey), this Patent discloses a method of clustering documents. Applicant agrees with the Examiner that, as part of this process, Tukey discloses the identification of an attractor for each cluster. However, since Sahami does not use attractors to form its clusters, Tukey and Sahami are technically incompatible, i.e. Tukey's cluster attractors are of no use to Sahami since Sahami's clusters are constructed as decision trees based on the influence of the "attributes" of the data. Hence, it is respectfully submitted that a combination of Tukey and Sahami could not lead a skilled person to Claim 1.

More generally, it is noted that clustering algorithms do exist that are based on identifying attractors and then grouping documents around the closest attractors. One example of this is the "K-means" described in paragraphs [0012] to [0015] of Sahami.

The key problem with this type of technique is how to identify good attractors in an efficient manner. This problem is addressed by Claim 1 of the present invention. It is noted that Tukey does not present any algorithms for identifying cluster attractors. Instead, Tukey refers to an earlier Patent, namely US 5,442,778 (Pedersen). Pedersen discloses two algorithms for identifying cluster attractors. However, both of these are based on agglomerative clustering which is very different from the method recited in Claim 1.

It is respectfully submitted, therefore, that Claim 1 is not obvious over the combined teachings of Sahami and Tukey. It is also submitted that independent Claims 13, 14, 15 and 18, being of corresponding scope to Claim 1, are also novel and non-obvious in light of Sahami and Tukey.

With regard to the Examiner's rejections of the dependant claims, Applicant replies as follows.

With regard to the Examiner's comments on claim 2, as described by Sahami, the K-means technique is based on centroid identification. This identification is totally different from the identification of attractors in the present invention. In K-means centroids are identified in an iterative process based on a distance measure (e.g., Euclidean or Hamming distance - see paragraphs [0012] and [0013] of Sahami). This is in contrast to features C, D and E of claim 1 recited above. In particular, because K-means does not employ the above-identified feature C of claim 1, it follows that K-means cannot disclose the features of claim 2. It is also noted that Sahami discloses K-means as prior art and does not actually use it itself.

With regard to the Examiner's comments on claim 3, in the teaching of Sahami conditional probability is used to evaluate the influence of an attribute given a cluster of data records.

This is in contrast to the use of conditional probability in claim 3 which is concerned with the occurrence of co-occurring terms in a document. As indicated previously, Sahami is not concerned with the terms (words) of a document, but is concerned only with the attributes of the data set.

With regard to the Examiner's comments on claim 4, the "indicators" referred to in claim 4 comprise probability distributions. The Dice and Jaccard coefficients are normalized mathematical functions, in this case word overlap counts, but they are not probability distributions.

With regard to the Examiner's comments on claim 5 and 6, in addition to the differences highlighted by the Applicant in relation to claim 1, Applicant also notes that, in Sahami, the phrase "a subset of the set of data" means set of data records that are members of same cluster. This is not the same as the subsets of terms (words) recited in claim 5. means a set of terms/words from a document corpus which meet a specific condition on

With regard to the Examiner's comments on claim 7 and 8, as indicated above, the Sahami teaches the use of entropy only to eliminate attributes that do not influence the process of database record clustering. This elimination is used only to improve the efficiency of Sahami's clustering process. Sahami does not disclose the use of entropy in the selection of cluster attractors, not least because Sahami does not use cluster attractors. Tukey does not disclose any method for attractor identification. Indeed he presents a reference on another patent, Pedersen, that discloses two such algorithms. However these algorithms are based on agglomerative clustering approach and are not based on entropy.

With regard to the Examiner's comments on claim 9, in contrast to claim 9, in Sahami the phrase "computes the frequency information" means evaluation of a number of database records that meet a specific condition. In Tukey the phrase "disjoint" relates to a special type of clustering ("hard" clustering) when any data record belongs to one and only one cluster. In claim 9, the phrase "subsets are disjoint" means that corresponding frequency ranges have an empty intersection. This is a very different concept to the one disclosed in Tukey.

With regard to the Examiner's comments on claim 10, in claim 10 the word "frequency" refers to the frequency of occurrence of terms (words). In Sahami ([0076]) this concept has a very different meaning, namely the number of records in a cluster that contain a given pair (attribute, attribute-value).

With regard to the Examiner's comments on claims 11 and 12, the same comments apply as were made in relation to claims 7 and 8.

With regard to the Examiner's comments on claim 16 and 17, as indicated above Sahami does not operate on the terms (words) of a document. Instead, its analysis is performed on "attributes" of the data. Hence, Sahami does not disclose the calculation of probability distributions of the occurrence of terms of each document. Tukey does not disclose the calculation of probability distributions of the occurrence of terms of each document or the comparison of these against probability distributions selected as cluster attractors since is uses agglomerative clustering techniques taught by Pedersen.

Accordingly, it is respectfully submitted that the dependent claims are novel and non-obvious when compared to Sahami and Tukey in their own right and not just because they are dependent on one or other of the independent claims.

The Applicant respectfully requests that a timely Notice of Allowance be issued in this case.

Respectfully submitted,

The Nath Law Group

By: ____________________________
Gary M. Nath
Registration Number 26,965
Jerald L. Meyer
Registration Number 41,194
Customer Number 20529

Date: May 27, 2008
THE NATH LAW GROUP
112 South West Street
Alexandria, VA 22314

GMN/JLM